Harnessing Feedback Region Proposals for Multi Object Tracking

ISSN 1751-8644
doi: 0000000000
www.ietdl.org

Aswathy P.[1]*, *Deepak Mishra*[1]

[1] *Department of Avionics, Indian Institute of Space Science and Technology, Trivandrum, Kerala, India*
* *E-mail: aswathyece2011@gmail.com*

Abstract: In tracking-by-detection, a major challenge in online multiple object tracking (MOT) is how to associate object detections on the new video frame with previously tracked objects. Two important aspects that directly influence the performance of MOT are quality of detection and accuracy in data association. We propose an efficient and unified MOT framework for improved object detection followed by enhanced object tracking. The object detection and tracking are considered as two independent functions in tracking-by-detection paradigm, where the detection responses at each video frame need to be reliably linked to form target trajectories. In this study, object detection accuracy has been increased by employing Faster R-CNN modified with the feedback region proposals from the tracker. Target association is performed by correlation filter based Siamese CNN model trained on reinforcement learning strategy, that finds the similarity score between the input image patches. An optical flow based motion model is employed to predict the next probable location of the targets from the tracker and these region proposals are fed back to classifier module of Faster R-CNN. Our extensive analysis on publicly available multiple object tracking benchmark datasets and comparison with the state-of-the-art tracking methods demonstrate the superior performance of proposed MOT framework.

1 Introduction

Multiple object tracking is the process of localizing multiple moving objects over time. The problem of tracking multiple objects in a video sequence poses several challenging tasks include estimation of time-varying number of objects, motion prediction of all objects, object re-identification, similar appearance and dealing long and short term occlusions. A common approach for solving multi-object visual tracking problem is tracking-by-detection. In tracking-by-detection paradigm of MOT: first an object detector is applied to each frame of the video to locate the objects of interest, while using data association algorithms, a unique identity is assigned to every detected object. These identities are linked across a sequence of frames to form object trajectories.

On-line and batch methods are two commonly adopted methods for trajectory extraction. On-line methods [1],[2],[3] use the current and previous frames detections to object state estimation at each time epoch. The batch techniques [4],[5],[6] require the observations from a batch of frames in advance to estimate the final object state. It is due to this, the batch methods are difficult to use reliably in real time applications. The proposed multiple object tracking-by-detection framework works in online fashion.

In MOT, generally object detector and tracker are considered as two independent modules where the detection responses need to be reliably linked to form target trajectories. However, the performance of the tracker heavily rely on the quality of detection results from the object detector. Recent advancements in deep-learning-based object detection systems [7],[8],[9],[10] have made MOT more robust. In the proposed MOT framework, the highly efficient Faster R-CNN (Region-based Convolutional Neutral Network) [9] is used as the person detector module. In most of the situations, object detectors does not involve temporal information. In the proposed method, feedback region proposals from the past object state estimation is given to the object detector. This introduction of feedback helps to reduce the missed detection in the video sequences and thus improves the overall MOT accuracy.

When the objects are detected by a detector, what matters most is to associate the current detections with the existing tracks using an efficient data association algorithm. If there is a missing or inaccurate detection, the target is prone to be lost. To alleviate such issues, the proposed MOT approach integrates the merits of single object tracking and data association methods in a unified framework. A single object tracker uses the detection in the first frame and updates the appearance model online to find the target in the subsequent frames. Data association method compute the similarity between the detections in the frame and tracklets from the previous frames. Significant advancements in MOT have been achieved after deploying deep features, in particular Siamese networks [11],[12],[13],[14], for reliable and robust data association. In Siamese based architecture, the input is a pair of image patches. The network learns a similarity metric between the patches and output a similarity score between them. In the proposed method, a correlation filter based Siamese CNN (CFNet) [12] is used in both roles; as a single object tracker and as a data association method. CFNet is trained for data association in a reinforcement learning fashion which benefits from both advantages of offline-learning and online-learning. The positive and negative training samples for learning are generated according to the online MOT tracking results with the supervision from ground truth trajectories.

In this paper, we introduce an efficient multiple object tracking framework with an improved detector and efficient data association algorithm. In order to handle the missed detections, feedback region proposals from the tracker is presented to the object detector. Data association problem is tackled by a re-identification Siamese CNN model (CFNet). The proposed method also benefit from the strengths of CFNet as online single object tracker, where appearance of object is updated for target to handle detection failures. In addition to this, the Siamese network is trained using a reinforcement learning strategy, which helps the CNN model to learn the context in MOT dataset. Comprehensive analysis are performed on publicly available MOT benchmark datasets, MOT15, MOT16 and MOT17. The comparative results against the state-of-art methods show that our method performs substantially better in terms of common metrics used in MOT literature.

The outline of the paper is as follows. In section 2, we review the background on the multiple object tracking, object detection methods and the Siamese CNN models. Section 3 describes the proposed tracking frame work in detail. The employed Faster R-CNN module with the feedback region proposals framework is briefly explained in section 3.1, while the data association methodology that incorporate the CFNet is discussed in section 3.2. The algorithm for the proposed multiple object tracking is summarized in section 3.4. Experimental evaluations and results are detailed in section 4. Finally we conclude

the paper with a brief discussion on possible future directions in section 5.

2 Related Works

Multiple Object Tracking: Despite the recent advances on multiple object tracking, it still remains a complex and difficult task in the crowded environments with frequent occlusions, similar appearance, false detections etc. Mainly the MOT methods can be classified into three; (i) the data association problem modelled as an optimization problem or graphs [15],[16], (ii) solve data association problem using an end-to-end neural network [17],[18], (iii) use MOT paradigm other than tracking-by-detection [19]. The first two categories gives solution with a tracking-by-detection approach, where detector and tracker exists as two independent functions. In our framework, these two modules co-exists as feedback from tracker is given to the detector in frame-to-frame. The third category is aims to search for novel and more simple MOT methods, but trade-off between performance and speed still needs improvement.

Object Detection: Most of the recent multiple object trackers follow tracking-by-detection approach which heavily depends on the quality of object detection. Initially, most of the object detectors relied on using handcrafted features. But the introduction of deep convolutional neural networks have demonstrated remarkable improvement in the performance of object detectors. In [20], the idea of selective search is employed for proposing probable object locations. Recent development in object detection are driven by the success of region based CNNs(R-CNN) [21]. Advances like SPPNet [7] and Fast R-CNN [8] have improved the detection performance with reduced running time. The Faster R-CNN detector [9] and further SDP detector [22] employ fully convolutional neural network for region proposals and classification without any handcrafted features. Then Redmon et al. [10] proposed You Only Look Once (YOLO) detector that bypass the need for a region proposal network.

Siamese Convolutional Neural Networks: In MOT, data association can be addressed using similarity learning between image patches. Siamese convolutional neural network is widely used for similarity measurements and the CNN architecture used in the model gives a better image feature representation. Bertinetto et al. [11] proposed a fully convolutional Siamese network to measure the similarity score between image patches, which is employed for object tracking. CFNet [12] is an asymmetric architecture that incorporates a correlation filter into the Siamese network. Other variants to Siamese convolutional neural network includes DSiam [23] that uses fast transfer motion to update the model, SINT [24] that make use of optical flow methods and SA-Siam [25] that utilizes the combination of original Siamese architecture. In the proposed MOT framework, Siamese network performs in two roles: as a single object tracker and as a similarity function.

3 Online Multiple Object Tracking Framework

We propose an online multiple object tracking framework that uses two well known CNN architectures tailored according to MOT framework, in particular we use Faster R-CNN [9] for person detection and correlation filter based Siamese network (CFNet) [12] for data association. To enhance the accuracy of the detection and to reduce missed detections, here we introduce feedback region proposals from the tracker to the detector, using the motion model of the target objects. In tracking-by-detection, the detections in the current frame is associated with the existing tracks using an efficient data association algorithm. In this study, we exploit the merits of both single object tracking and data association to maintain target identities in a unified MOT framework. The proposed MOT method benefit from the strengths of CFNet as online single object tracker and similarity based data association algorithm. In the following sections, a detailed description on the proposed online MOT algorithm is given.

3.1 Faster R-CNN Detector with Feedback Region Proposals

To perform person detection, we employ Faster R-CNN architecture in our framework. A block diagram representation of the Faster R-CNN object detector with feedback region proposals is given in Fig. 1. There are two main stages in Faster R-CNN detector, a region proposal network (RPN) and a classifier-regressor module. The RPN generates a multitude of region proposals for each probable object of interest in the input image. The prerequisite step for RPN is feature extraction using convolution neural network. To generate good quality feature maps, we use ResNet-101 as our backbone CNN network. RoI (Region of Interest) pooling method is used to extract the feature maps for the proposals in uniform scale and size. RoIs are then proceeded to the classification and bounding box regression modules. The classifier assigns a class score for each RoIs and regression module realigned the bounding box to fit with the object. The final set of detections are obtained after employing a non-maximum-suppression (NMS) step to the bounding boxes. In this study, we are interested only in the person category after detection.

Generally, in multiple object tracking, detection and data association are considered as two independent functions. The detector does not require any temporal information of objects to perform its task. Here, we investigate the performance enhancement of an object detector provided with temporal information about the past target trajectories. The probable locations of the existing targets (tracked and lost targets in the previous frame I^{f-1}) in the current frame, I^f are predicted using its motion model and are feedbacked to the detector as region proposals. Let B^{f-1} represents the set of target bounding boxes at the frame I^{f-1}, which includes the tracked and lost targets.

$$\begin{aligned} B^{f-1} &= \{B_T^{f-1}, B_L^{f-1}\}, \\ B_T^{f-1} = \{b_j^{f-1}\}_{j=1}^{N_T};\ B_L^{f-1} &= \{b_k^{f-1}\}_{k=1}^{N_L}, \\ \text{where } b_i &= \{x, y, w, h\}. \end{aligned} \tag{1}$$

In the equation (1), N_T and N_L are the number of tracked and lost targets, (x, y) are the centre coordinates of the target, and (w, h) are the width and height of the target. In order to get the new target location, we compute an optical flow from densely and uniformly sampled points inside the current target template to the new video frame. Specifically, given the current target position, $p = (x, y)$, we find its corresponding location $p^* = p + u = (x + u_x, y + u_y)$ in the new frame using the iterative Lucas-Kanade method with pyramids [26], where $u = (u_x, u_y)$ is the optical flow at point p. We can predict the new bounding box for the target with center p^* and size same as previous box (w,h), which is treated as the new region proposal for that target. The set of all the predicted region proposals of the current targets B^{*f-1}, tracked and lost ones, are given back to the Faster R-CNN detector module. Along with the proposals from RPN module of the detector, the feedback region proposals from the tracker is provided to the RoI pooling. The remaining flow of the Faster R-CNN is unchanged.

In the evaluation, the two measures that affect the performance of MOT algorithm are identity switches and fragmentation issue. The number of times a particular target changes its identity is measured by identity switches. When the object is not detected in some frames, then fragmented trajectories are generated. These two issues occur mainly due to the missed detections in the video frames. The feedback proposals from the previous tracks in the proposed MOT algorithm helps to reduce the missed detections in each frame and thereby reduces the identity switches of the target and fragmented trajectories. This in effect improve the overall detection and tracking accuracy.

3.2 MOT - Data Association Methodology

In this section, we present the multiple object tracking framework where we incorporate the correlation filter based Siamese network (CFNet) to solve the data association problem. For each video frame,

Fig. 1: Faster R-CNN Object Detector with feedback region proposals. For each input feame I^f, Faster R-CNN detector outputs detection bounding boxes considering all the region proposals from RPN module and feedback region proposals from the past trajectories.

Faster R-CNN provide person detection regions. Data association algorithm identify a correspondence between the new object detections and pre-existing tracks. Here we integrates the merits of single object tracking and CNN based similarity metric for data association. Siamese CNN performs better as both single object tracker and similarity network.

Fig. 2: Target state transition in the proposed MOT framework.

In this MOT framework, we adopt the state transitions of the target explained in [27] with some modifications. A target in the video can go through 4 life stages: initialized, tracked, lost, inactive. Fig. 2 illustrates the state transitions of the targets between the four stages. The object trajectory is initialized when that object appeared in the video frames for the first time. In the first frame, all the detections from the detector is considered as tracked targets and for each detection a new trajectory is initialized in the trajectory list. Now, the single object tracker has to take decision whether to keep each target in tracked state or transfer it into lost state. The state of the target is set as tracked till it is not occluded or is not out of camera's field of view. Otherwise, the target is regarded as lost. This decision making is related to the tracking score and consistency of tracking result with the object detections. Once the object is entered into the lost state, data association algorithm tries to find out a match for the lost targets with the detections, that are not covered by any tracked target. If the similarity function could find a matched detection for the lost target, the state is updated as tracked and tracking process resumes for the same. If the target stays in the lost state for a long time, it is considered that the object entered into inactive state and we terminate the trajectory corresponds to that object

In the proposed method, tracking problem is addressed by using a correlation filter based siamese CNN (CFNet) model that predict whether two image patches belong to the same trajectory or not. The functions of the CFNet here are two fold:

- If the target is in the tracked state, CFNet works as a single object tracker, that find out the new target location.
- If the object is occluded, ie, in lost state, CFNet acts as a patch similarity function, that gives similarity score between lost target template and detections from detector module that are not associated with tracked objects.

The pre-trained CFNet architecture with learned weights is adapted and is retrained with MOT benchmark dataset in a reinforcement learning fashion for our purpose.

3.2.1 CFNet as Single Object Tracker: Visual object tracking algorithms based on Siamese CNN architecture formulate tracking as a template matching problem [11],[24],[12],[14]. The network structure has two identical CNN branches that share the kernel weights. One branch extracts the feature maps of the target image patch and the other one of the search image patches which contain the candidate objects. The search area is chosen with center same as the previous target location and size 2.5 times larger than the target image. A number of candidate image patches with same size as target is chosen within the search area. Then we obtain a similarity score map by the cross-correlation between the convolutional feature maps of target and candidate patches. In object tracking, the goal is to find the new target location and is obtained from the most similar candidate image patch.

Let x_T represents the target patch and x_c represents the candidate patch. These inputs are processed by the CNN, ϕ_p, where p is the learnable parameters. Then the feature maps $\phi_p(x_T)$ and $\phi_p(x_c)$ are cross-correlated as given by,

$$\Psi_p(x_T, x_c) = \phi_p(x_T) \star \phi_p(x_c) \tag{2}$$

The correlation filter based Siamese network (CFNet) [12] incorporates two additional layers within the baseline Siamese network [11], correlation filter and crop layers, which makes it is more shallower and faster without accuracy drop. The correlation filter layer inserted between the CNN with target patch and cross-correlation module estimates the discriminative features of the target patches. Then the modifications in (2) can be formulated as,

$$\Psi_{p,\alpha,\beta}(x_T, x_c) = \alpha\omega_c(\phi_p(x_T)) \star \phi_p(x_c) + \beta, \tag{3}$$

where $\omega_c(.)$ represents the correlation filter layer that learns during training, by solving a ridge regression problem in Fourier domain. In this equation, α and β are scale and bias parameters. The maximum in similarity score map, corresponds to the new target location.

$$\hat{x_T} = \arg\max_{x_c^i} \Psi_{p,\alpha,\beta}(x_T, x_c^i) \tag{4}$$

3.2.2 CFNet for patch similarity: When the target is in the lost state, data association algorithm has to decide whether to keep the target in lost state, move in to tracked state or terminate the target trajectory (inactive state). In order to move from lost state to the tracked state, any of the detections from the person detector needs to be associated with lost target. A Siamese network can also be used as a similarity function that check the pair wise similarity between the lost target patch and the detections. Fig. 3 demonstrates the CFNet architecture as similarity metric in data association. Let x_L represents the lost target image patch and $D^f = \{d_i^f\}_{i=1}^{N_f}$ are the set of detections given in the current frame, I^f. CFNet outputs N_f similarity score maps, each corresponding to the match score between the detection and the lost target. If the maximum similarity score s_m is above the threshold T_s, then the detection and lost target pair is considered for data association. Hungarian algorithm is employed to assign the detections to the lost targets. If a detection is get associate with the lost target, the it is transferred to the tracked state and the detection corresponds to the maximum score, d_m^f is updated as the target image, $\hat{x_L}$.

$$d_m^f = \arg\max_{d_i^f} \Psi_{p,\alpha,\beta}(x_L, d_i^f); i = 1 : N_f \tag{5}$$

$$s_m = \max(\Psi_{p,\alpha,\beta}(x_L, d_m^f)) \tag{6}$$

$$state = \begin{cases} tracked, & \text{if } s_m \geq T_s \\ lost, & \text{otherwise} \end{cases} \tag{7}$$

$$\hat{x_L} = d_m^f; \text{ if state=tracked} \tag{8}$$

3.3 Training CFNet - Reinforcement Learning

In this study, the CFNet architecture that trained offline [12] is used as a back bone Siamese CNN. Within the context of multiple object tracking, the weight parameters of the network are then retrained using a reinforcement learning strategy. The main advantage of reinforcement learning is that it does not require any labelled data for

Fig. 3: CFNet for patch similarity in lost state.The CFNet cross-correlate the image patches presented and obtain the similarity score maps. The detection with maximum similarity is considered to associate with lost target.

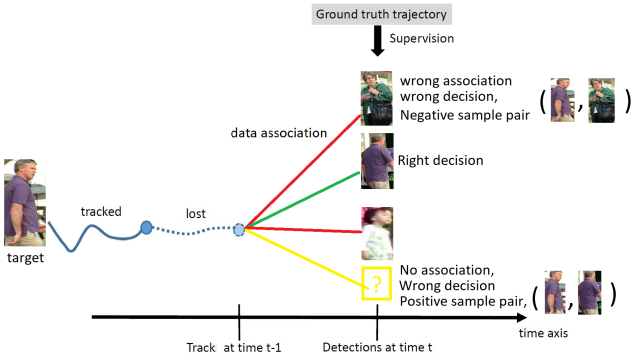

Fig. 4: Reinforcement Learning Method. Positive and negative training samples are generated based on the data association system output.

training. It learns from the delayed label called reward from its system environment. Reward is a scalar value that indicates whether the decision of that system is right or wrong. Hence, the goal of learning method is to take actions in order to maximize reward.

The Siamese network learns the similarity function during training from the positive and negative image pairs. With the supervision from the ground truth tracks, training samples are generated from the online tracking results. The approach we used to generate the training examples is shown in the Fig. 4. Let x_L denotes the lost target and $D^f = \{d_i^f\}_{i=1}^{N_f}$ represents the set of all the detections in the present frame. Based on the similarity score for data association, we assign a label, $z \in (-1, 1)$ to the image pairs, (x_L, d_i^f) that indicate whether the lost target is associated (z=1) or not (z=-1) to the detections. During training process, a reward function, $R_L(x_L, z)$ is evaluated based on the ground truth track and the decision from the data association module. To formulate the reward function, we need another function variable $y(x_L) \in (-1, 1)$ whose value depends on the ground truth trajectory corresponding to target, x_L.

$$y(x_L) = \begin{cases} 1, & \text{if } \max(overlap\_ratio(x_L^{GT}, d_i^f)) \geq T_o \\ -1, & \text{otherwise} \end{cases}, \quad (9)$$

where x_L^{GT} is the ground truth track for the lost target and T_o is the overlap threshold. The equation (9) here means that, $y = +1$ if there is a match for x_L in the given detections and $y = -1$ if there is no association possible, according to the ground truth.

$$z_m = \max_{i=1}^{M} z(x_L, d_i^f) \quad (10)$$

If the data association algorithm assigns a detection d_m^f with the target, then the value of $z_m = +1$ and if the decision is not to associate the target with any of the detections, $z_m = -1$.

$$R_L(x_L, z) = y(x_L) * z_m, \quad (11)$$

The Table 1 tabulates the reward for different decisions z on x_L by the data association rule. It is clear from the table that reward, $R = +1$ if the data association took a right decision and $R = -1$ if the decision is wrong. The data association system, CFNet need to get updated only when it makes a mistake in decision making. i.e., association estimation, z_m takes different action as desired by the ground truth trajectory. There are two cases for which the reward

Table 1 Reward for decision making in data association in lost state.

$y(x_L)$	z_m	$R(x_L, z)$
1	1	1
-1	1	-1
1	-1	-1
-1	-1	1

is negative. In the first case, the target is linked ($z_m = 1$) to an object detection d_m^f (refer (5)-(8)), which is incorrect according to the ground truth ($y = -1$). Then the image pair (x_L, d_m^f) is added to the training database as a negative sample. In the second case, the decision is not to link ($z_m = -1$) with any of the detections given. But the target detection, d_k^f is correctly included in the detections ($y = 1$) and association algorithm missed the right association. Then the image pair (x_L, d_k^f) is included to the training database as a positive sample, where

$$d_k^f = \arg\max_{d_i^f}(overlap\_ratio(x_L^{GT}, d_i^f)). \quad (12)$$

In the online training process, we start with the CFNet model pre-trained on ImageNet dataset. Using the new training samples, the CFNet model is retrained and updated the parameters p, by minimizing the logistic loss over the new training data.

$$\arg\min_{p} \sum_{j} \mathcal{L}(\Psi_{p,\alpha,\beta}(x_j, d_j), y(x_j)) \quad (13)$$

The trained CFNet model in the context of multiple object tracking performs better under the challenging conditions and provides better the accuracy in tracking.

3.4 Multiple Object Tracking Algorithm

After training the Faster R-CNN with MOT-17 dataset and CFNet using the Reinforcement learning method, we apply these trained architectures for multiple object tracking. The proposed MOT algorithm is summarized in Algorithm 1. Given an image sequence, the goal of MOT problem is to estimate the optimal sequential states of all the targets, ie, the trajectory of each target. For each input frame, the Faster R-CNN detector outputs a set of person detection, D^f. In the data association part, targets in the tracked state gets higher priority and they are processed first. The CFNet as a single object tracker determines whether the target should stay in tracked state or moved to the lost state. Then, for the lost targets, the CFNet acts as a similarity function computes the pairwise similarity score with the object detections from the detector. Hungarian algorithm is then employed to associate the detections to the lost targets based on the similarity score. The targets that are linked with the detections are reassigned as tracked targets. Finally, the detection that are not associated with any of the tracked targets are considered as new targets and a trajectory is initialized for each new detection. Here, to avoid the covered detections, non-maximum suppression based on bounding box overlap is applied. To improve the detection performance, we incorporate feedback region proposals that provide a temporal information about the previous trajectories, to the detector.

4 Results

This section presents the experimental results of the proposed multiple object tracker on benchmark datasets focusing on person tracking to validate the efficiency and the tracking performance. To

Algorithm 1 Online Multiple Object Tracking with Faster R-CNN and CFNet

Input: Video Sequence as an ordered list of image frames, $V = \{I^f | f = 1, 2, \ldots, F\}$

Output: Set of object trajectories, $\mathcal{T} = \{\tau_i\}_{i=1}^{N}$, with $\tau_i = \{b_i^{f_s}, \ldots, b_i^{f_e}\}$, as a list of ordered target bounding boxes, where f_s and f_e are first and last frame in which target i exists, $b_i^{f_j} = (x, y, w, h)$

Initialization: $\mathcal{T} \leftarrow \emptyset$
Trained CNN models: Faster R-CNN with feedback region proposals retrained on MOT dataset as multiple object detector and CFNet trained using reinforcement learning as data association metric.
for Video frame I^f in V **do**
  Person detections from Faster R-CNN, $D^f = \{d_j^f\}_{j=1}^{N_f}$
  if $(f == 1)$ **then**
    Initialize new trajectory τ_i^f for each detection,
    state==tracked;
  else
    for each tracked target $\tau_i^{(f-1)} \in \mathcal{T}$ **do**
      CFNet as single object tracker; find the new target location and similarity score, s_m.
      if s_m<T_s **then**
        state==lost;
      end if
    end for
    for each lost targets, $\tau_i^{(f-1)} \in \mathcal{T}$ **do**
      CFNet as patch similarity function;
      for each detection d_k^f not covered by tracked objects **do**
        Obtain the similarity score map with lost target
        Compute the maximum similarity score, s_m
        if $s_m > T_s$ **then**
          detection considered for association with lost target
        end if
      end for
      Hungarian algorithm employed to assign detection to the lost targets.
      if lost $\tau_i^{(f-1)}$ assigned to detection d_m^f **then**
        state ==tracked;
      end if
    end for
    for each detection d_j^f not covered by tracked and lost targets **do**
      Initialize a trajectory τ_i^f.
      state==tracked.
    end for
  end if
  for each $\tau_i^f \in \mathcal{T}$ **do**
    Predict the next location v using Lucas Kanade optical flow method.
    Feedback region proposals, R=set of bounding boxes with center v and size same as τ_i.
    (Faster R-CNN classifier module cascade R along with its RPN module output).
  end for
end for
return set of trajectories of the objects, $\mathcal{T}$.

obtain comparable results with the state-of-the-art trackers, we evaluated our tracker framework on MOT challenge dataset, a standard reference when addressing MOT problems.

MOT Challenge: MOT challenge dataset is a centralized benchmark dataset to test the MOT tracking methods that includes several challenging tracking sequences with different characteristics such as object density, frame rate, occlusions,illuminations etc.. Mainly, there are three separate tracking sequence set published by MOT challenge, 2DMOT2015, MOT16 and MOT17. Each of the benchmark dataset includes separate video sequences for training and testing of the tracker. Training sequences are provided with public object detections from object detectors and the groundtruth detections, whereas testing sequences only include object detections. The MOT17 dataset contains 14 challenging sequences of which seven are used for training and seven for testing the tracker. The sequences are provided with three sets of detection from DPM, Faster R-CNN and SDP object detectors. The benchmark sequences included in MOT16 are same as that of MOT17 with only DPM detection. The benchmark dataset 2DMOT2015 includes total of 22 sequences each of which provided with ACF detections.

There are several metrics for the quantitative evaluation of multiple object tracking that measure different aspects of tracker efficiency. The two important parameter in MOT17 challenge that measure the object coverage and identity consistency are MOTA (Multiple Object Tracking Accuracy) and IDF1 score. MT(Mostly Tracked) and ML(Mostly Lost) are another two parameters that indicate the percentage of groundtruth objects whose trajectories are covered by the tracking output. FP and FN represents respectively, the total number of false positives and false negatives. IDSw The number of times a particular target changes its identity is measured by identity switches (IDSw). When the object is not detected in some frames, then fragmented trajectories are generated (Frag).

4.1 Analysis on Validation Dataset

Table 2 Training and Testing sequences for validation on the MOT Benchmark.

Training	Validation
2D MOT 15	
TUD-Stadtmitte	TUD-Campus
ETH-Bahnhof	ETH-Sunnyday, ETH-Pedcross2
ADL-Rundle-6	ADL-Rundle-8, Venice-2
KITTI-13	KITTI-17
MOT17	
MOT17-02	MOT17-04
MOT17-05	MOT17-09
MOT17-10	MOT17-11, MOT17-13

In our MOT frame work, detector and tracker works together hence are not considered as separate units. The Faster R-CNN detections in each frame depends on the past tracking output, as detector takes feedback region proposals. Therefore the object detections provided in MOT challenge database are not directly used in the analysis. The main contributions in this study are (i) proposed a temporal information to the detector in the form of feedback region proposals to reduce the missed detections, (ii) introduced an online reinforcement learning method to train the Siamese network based similarity function in data association and (iii) integrated the benefits of single object tracker and similarity metric for data association in a unified MOT framework. We conducted our experiments to validate the importance of each contributions in the proposed MOT algorithm. Since the object detection annotations of the MOT test dataset are not released, we use the MOT training sequences to conduct analysis about our framework. The training data set is divided into training and validation sequences. The splitting of the sequences is shown in table 2.

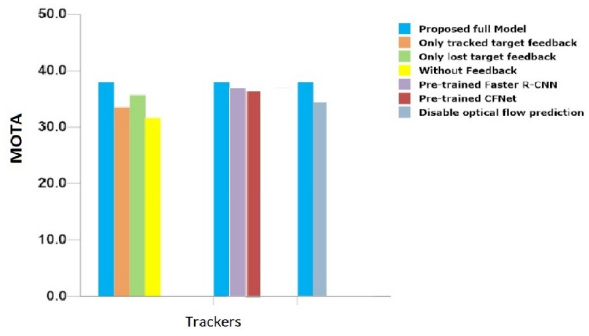

Fig. 5: Analysis of the proposed MOT framework on the validation set (2D MOT 2015) with different components.

4.1.1 Contribution of Different Components: We investigate the contribution of different components in our framework by disabling one component at a time and then examining the performance drop in terms of MOTA on the validation set. Fig. 5 shows the significance of each components validating using 2DMOT15 benchmark dataset. The table 3 also presents the importance of each components evaluated on the MOT17 validation dataset. In this fig. 5, first set of bars presents the importance of feedback region proposals from the tracker to the detector. It is evident that the feedback improves the accuracy of the MOT framework. One of the challenging factor that affect the MOTA value in MOT evaluation is identity switches (IDSw), a measure that indicates the number of times the target changes its identity in the whole tracking process. Fragmentation is another factor that affect the performance of our proposed system. Fragmented trajectories are formed when identity switches does not occur, but the detector missed the target detection. The solution for these two problems is to reduce the number of missed detections. This can be done by improving the performance of the person detector. The feedback region proposals can be viewed as a reference given to the detector on the probable locations of existing targets. This helps the detector to reduce the missed detections and improve its efficiency. In the proposed full model MOT, we sent back the region proposal prediction for both tracked and lost targets. For the detailed analysis on this feedback proposals, we consider three different situations on feedback. (i) Only feedback tracked targets proposals (no lost targets), (ii) only feedback lost targets proposals and (iii) no feedback proposals. It is obvious from the results that with the feedback region proposals the performance of our tracking framework is improved. In addition to that, it is interested to note the accuracy difference with the region proposals with any one of the target state, tracked or lost. For the target in the tracked state, the appearance model of the tracked target is getting updated in each frame. Also, in the proposed MOT framework, a single object tracker is used to track the tracked targets. Therefore, chances are less for a tracked target to be in a lost category even if the detector missed the target detection. But in the case of lost targets, the position or the appearance model is not updated and the data association is completely rely on the detections from the detector. Therefore, if the detector missed the correct target detection, the target will be continuing on its lost state and the tracking accuracy reduces. It is clear from the fig. 5 and the table 3 that the proposed feedback region proposals reduce the missed detections and improves the overall performance of the MOT system.

In the fig. 5, second set indicates the significance of online training on the pre-trained CNN networks used in the proposed tracker. We employed a retrained Faster R-CNN with MOT dataset as the person detector and Siamese CFNet model trained online using reinforcement learning method on MOT dataset as a tracker. The Faster R-CNN object detector with ResNet-101 pre-trained on PASCAL-VOC and COCO training set is retrained on the MOT17 person detection dataset. As discussed in section 3.3, the pre-trained CFNet is trained online using reinforcement learning method for data association. It is clear from both the fig. 5 and the table 3 that the detector and tracker CNN networks improve its accuracy after trained on MOT benchmark dataset.

The third set in fig. 5 shows the relevance of motion model of the target to predict the feedback region proposals. The motion model based on Lucas-Kanade optical flow method with pyramids is employed here to predict the new location of the tracked or the lost targets. To study the contribution of the motion model, current target location is fed back to the detector as the region proposals instead of the predicted location from motion model. The results shows an accuracy drop without optical flow based motion prediction. Since the target in the tracked state is moving slowly from one frame to the next (usually verified from high frame rates), the regression module in the Faster R-CNN detector is able to refine the proposal bounding box of that slightly shifted target. But in case of lost targets, the position of the target is not updated for all the lost frames. Therefore, the regression module could not find the refined location of the particular region proposal. If the detector's own proposals doesn't include the lost target, then it cause a missed detection. By using the optical flow based motion prediction, most of the time we could predict the location of the occluded targets and could avoid these missed detections and thus improve the tracking performance.

4.2 Evaluation on Test Dataset

The propose MOT framework is evaluated on MOT test dataset of both MOT17 and 2DMOT2015. Some of the recent and better performing multiple object trackers are selected for the comparison with our approach. The state-of-the-art trackers evaluated with public detections provided by MOT challenge For the fair comparison with these trackers, the proposed MOT framework is slightly modified to perform all the test data evaluation with the given precomputed detections. Here, we are not using our Faster-RCNN detector to find a new detection and new tracks are initialized only from the frame to frame detections provided with MOT dataset. To deal with the proposed feedback region proposals, only the classifier and the regression modules of the Faster R-CNN are used. We forward the MOT detections along with the detections computed from the feedback region proposals to the MOT tracker part. To avoid multiple detection entries for the same object, a non-maximum suppression that based on bounding box overlap is employed before the tracker module.

The multiple object tracker trained on MOT training sequences, is then tested on the MOT17 and 2DMOT2015 testing dataset. Our experimental results are then submitted to the MOT challenge website for evaluation. Table 4 summarises the tracking performance of our proposed tracker on MOT benchmark, where we compare it with other tracking methods. The contribution of various proposed components such as feedback region proposals, online training based on reinforcement learning and motion prediction model are also given in table 4. It is clear from the results that the introduction of feedback region proposals model helps to reduce the missed detections. Thus, the number of identity switches (IDSw) got reduced and MOTA improved. The experiment results shows that, in comparison with other MOT methods our MOT framework achieves competitive tracking performance.

5 Conclusion

In this study, we developed a unified multiple object tracking framework with an efficient object detection module and an accurate data association method. The Faster R-CNN person detector with feedback region proposals from the tracker reduces the missed detections and provides better object detections that in turn help to improve the data association accuracy. The data association algorithm designed here exploits the strengths of correlation filter based Siamese CNN network (CFNet) as a single object tracker and a similarity metric. Furthermore, we proposed a reinforcement learning based training method to train the Siamese network in an online fashion. The complete multiple object tracking system is trained and evaluated over the benchmark MOT Challenge datasets. When comparing with the state-of-the-art trackers, it is observed that the proposed MOT

Table 3 Analysis of the proposed MOT framework on the validation set (MOT17) with different components.

Tracker	MOTA	IDF1	MT	ML	FP	FN	IDSw	Frag.
Proposed MOT	59.8	52.2	20.4	28.9	18122	201,011	1857	2648
Proposed MOT without feedback proposals	56.7	49.8	16.2	33.2	23158	263,284	2486	3915
Proposed MOT without Online Training	58.1	51.0	18.9	31.5	22189	238,138	2128	2831
Proposed MOT without motion prediction	57.9	51.4	19.7	30.3	22946	213,242	2291	2821

algorithm performs better in terms of MOT evaluation metrics. The evaluation results on the validation dataset also shows the relevance of each proposed components in the MOT framework. Moreover, our MOT framework is general to be associated with different methods employed in person detection, single object tracking and image similarity matching.

6 References

1 Breitenstein, M.D., Reichlin, F., Leibe, B., et al.: 'Online multiperson tracking-by-detection from a single, uncalibrated camera', IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33, (9), pp. 1820âĂŞ1833

2 Zhang, J., Presti, L., Sclaroff, S.: 'Online multi-person tracking by tracker hierarchy,' Proc. IEEE Int. Conf. Advanced Video Signal-Based Surveillance, Beijing, China, Sept. 2012, pp. 379âĂŞ385

3 Yoon,J.H., Lee, C.R., Yang, M.H., et al.: 'Online multi-object tracking via structural constraint event aggregation', Proc. IEEE Conf. Computer Vision and Pattern Recognition, Las Vegas, NV, USA, June 2016, pp. 1392-1400.

4 Chen, L., Peng, X., Ren, M.: 'Recurrent metric networks and batch multiple hypothesis for multi-object tracking', IEEE Access, 2018, 7, pp. 3093âĂŞ3105

5 Henriques, J.F., Caseiro, R., Batista, J.: 'Globally optimal solution to multi-object tracking with merged measurements', Proc. IEEE Int. Conf. Computer Vision, Barcelona, Spain, Nov. 2011, pp. 2470âĂŞ2477

6 Milan, A., Roth, S., Schindler, K.: 'Continuous energy minimization for multitarget tracking', IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36, (1), pp. 58âĂŞ72

7 He, K., Zhang, X., Ren S., et al.: 'Spatial pyramid pooling in deep convolutional networks for visual recognition', IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37 , (9) , pp. 1904 - 1916

8 Girshick, R.: 'Fast R-CNN', IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, Dec. 2015,

9 Ren, S., He, K., Girshick, R., et al.: 'Faster R-CNN: Towards realtime object detection with region proposal networks', Advances in Neural Information Processing Systems, 2015, pp. 91-99

10 Redmon, J., Divvala, S., Girshick, R., et al.: 'You only look once: Unified, real-time object detection', Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2016, pp. 779-788

11 Bertinetto, L., Valmadre, J., Henriques, J.F., et al.: 'Fully-convolutional siamese networks for object tracking', European conference on computer vision, Springer, 2016, pp. 850âĂŞ865

12 Valmadre, J., Bertinetto, L., Henriques, J.F., et al.: 'End-to-end representation learning for correlation filter based tracking', Proc. IEEE Conf. Computer Vision and Pattern Recognition, Honolulu, HI, USA, July 2017, pp.5000-5008

13 Dong, X., Shen, J.: 'Triplet loss in Siamese network for object tracking', Proc. European Conf. on Computer Vision (ECCV), 2018, pp. 459âĂŞ474

14 Li, B., Yan, J., Wu, W., et al.: 'High performance visual tracking with Siamese region proposal network', Proc. IEEE Conf. Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, June 2018, pp. 8971âĂŞ8980

15 Wen, L., Li, W., Yan, J., et al.: 'Multiple target tracking based on undirected hierarchical relation hypergraph', Proc. IEEE Conf. Computer Vision and Pattern Recognition, Columbus, OH, USA, June 2014, pp. 1282-1289

16 Kim, C., Li, F., Ciptadi, A., et al.: 'Multiple hypothesis tracking revisited', International Conference on Computer Vision, Santiago, Chile, December 2015, pp. 4696-4704

17 Sun, S., Akhtar, N., Song, H., et al.: 'Deep affinity network for multiple object tracking', IEEE transactions on pattern analysis and machine intelligence, 2019

18 Zhu, J., Yang, H., Liu, N., et al.: 'Online multi-object tracking with dual matching attention networks', International Conference on Computer Vision, 2018

19 Bergmann, P., Meinhardt, T., Leal-Taixe, L.: 'Tracking without bells and whistles', arXiv preprint arXiv:1903.05625, 2019

20 Uijlings, J.R.R., Van De Sande, K.E.A., Gevers, T., et al.: 'Selective search for object recognition', International Journal of Computer Vision. 2013, 104, (2), pp. 154-171

21 Girshick, R., Donahue, J., Darrell, T.,et al.: 'Rich feature hierarchies for accurate object detection and semantic segmentation', Proc. IEEE Conf. Computer Vision and Pattern Recognition, Columbus, OH, USA, June 2014, pp. 580-587

22 Yang, F., Choi, W., Lin, Y.: 'Exploit all the layers: Fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers', Proc. IEEE Conf. Computer Vision and Pattern Recognition, Las Vegas, NV, USA, June 2016, pp. 2129-2137

23 Guo, Q., Feng, W., Zhou, C., et al.: 'Learning dynamic Siamese network for visual object tracking', Proc. IEEE Int. Conf. Computer Vision, Venice, Italy, Oct. 2017, pp.1781–1789

24 Tao, R., Gavves, E., Smeulders, A.W.: 'Siamese instance search for tracking', Proc. IEEE conf. computer vision and pattern recognition, Las Vegas, NV, USA, June 2016, pp. 1420–1429

25 He, A., Luo, C., Tian, X., et al.: 'A twofold siamese network for real-time object tracking', Proc. IEEE Conf. Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, June 2018, pp. 4834-4843

26 Bouguet, J.Y: 'Pyramidal implementation of the affine Lucas Kanade feature tracker description of the algorithm', Intel Corporation, 2001, 5, pp. 1âĂŞ10

27 Xiang, Y., Alahi, A., Savarese, S.: 'Learning to track: Online multi-object tracking by decision making',International Conference on Computer Vision, Santiago, Chile, December 2015, pp. 4705âĂŞ4713

28 Yoon, Y., Kim, D., Yoon, K.,et al.: 'Online Multiple Pedestrian Tracking using Deep Temporal Appearance Matching Association', arXiv:1907.00831, 2019

29 Chu, P., Ling, H.: 'FAMNet: Joint Learning of Feature, Affinity and Multi-dimensional Assignment for Online Multiple Object Tracking', International Conference on Computer Vision, 2019

30 Keuper, M., Tang, S., Andres, B.: 'Motion Segmentation and Multiple Object Tracking by Correlation Co-Clustering', IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018.

31 Chu, P., Fan, H., Tan, C., et al.: 'Online Multi-Object Tracking with Instance-Aware Tracker and Dynamic Model Refreshment', Winter Conference on Applications of Computer Vision,Waikoloa Village, HI, USA, January 2019, pp. 161-170

32 Sadeghian, A., Alahi, A., Savarese, S.,: 'Tracking The Untrackable: Learning To Track Multiple Cues with Long-Term Dependencies', International Conference on Computer Vision, Venice, Italy, October 2017, pp. 300-311

33 Chu, Q., Ouyang, W., Li, H., et al.: 'Online Multi-object Tracking Using CNN-Based Single Object Tracker with Spatial-Temporal Attention Mechanism', International Conference on Computer Vision, Venice, Italy, October 2017, pp. 4846-4855

Table 4 Analysis of the proposed MOT framework on the test dataset (MOT17 and 2DMOT2015) with state-of-the-art-trackers.

MOT17								
Tracker	MOTA	IDF1	MT	ML	FP	FN	IDSw	Frag.
Proposed complete MOT	51.8	51.2	19.7	31.9	19122	240,011	1927	2943
Proposed MOT without feedback proposals	48.5	49.7	15.2	35.2	22168	263,284	3286	3879
Proposed MOT without Online Training	50.9	50.1	17.3	32.5	20199	239,138	2198	2954
Proposed MOT without motion prediction	50.7	50.4	18.9	33.3	20945	243,242	2891	3123
DMAN [18]	48.2	55.7	19.3	38.3	26218	263,608	2194	5378
DEEPTAMA [28]	50.3	53.5	19.2	37.5	25479	252,996	2192	3978
FAMNet [29]	52.0	48.7	19.1	33.4	14138	253,616	3072	5318
jCC [30]	51.2	54.5	20.9	37.0	25937	247,822	1802	2984
2DMOT2015								
Proposed complete MOT	39.8	48.2	22.1	28.9	4122	20011	421	928
Proposed MOT without feedback proposals	36.7	47.2	19.2	35.3	5168	23284	786	1242
Proposed MOT without Online Training	33.2	49.0	21.1	34.5	7199	28138	387	984
Proposed MOT without motion prediction	34.9	48.2	22.2	34.2	6945	23242	589	1134
KCF [31]	38.9	44.5	16.6	31.5	7321	29501	720	1440
AMIR15 [32]	37.6	46.0	15.8	26.8	7933	29397	1026	2047
JointMC [30]	35.6	45.1	23.2	39.3	10580	28508	457	969
AM [33]	34.3	48.3	11.4	43.4	5154	34848	348	1463